MAS223 Statistical Modelling and Inference
Examples

Chapter 1

Example 1: *Sample spaces and random variables.*

Let S be the sample space for the experiment of tossing two coins; i.e.

$$S = \{HH, HT, TH, TT\}.$$

Define the random variables

- X to be the number of heads seen,
- Y to be equal to 5 if we see both a head and a tail, and 0 otherwise.

Element of S	Value of X	Value of Y
HH	2	0
HT	1	5
TH	1	5
TT	0	0

Example 2: *Discrete random variables*

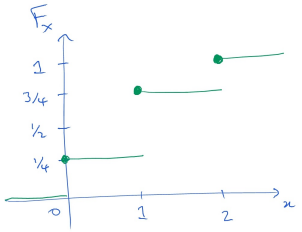
The random variables X and Y from Example 1 are both discrete random variables.

> *Calculate* $\mathbb{P}[X \leq 1]$.

If $X \leq 1$ then either $X = 0$ or $X = 1$. We have $\mathbb{P}[X = 0] + \mathbb{P}[X = 1] = \frac{3}{4}$.

> *Sketch the distribution function of* X.

A sketch of its distribution function looks like:



Example 3: *Continuous random variables*

> *Recall (from MAS113) that an exponential random variable with parameter* $\lambda > 0$ *has probability density function*

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Calculate $\mathbb{P}[1 \leq X \leq 2]$, *and find the distribution function* $F_X(x)$.

We can calculate

$$\mathbb{P}[1 \le X \le 2] = \int_1^2 f_X(x)\,dx = \int_1^2 \lambda e^{-\lambda x}\,dx = \left[-e^{-\lambda x}\right]_{x=1}^2 = e^{-\lambda} - e^{-2\lambda}.$$

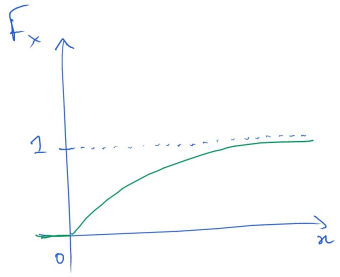
To find the distribution function, note that for $x \le 0$ we have $\mathbb{P}[X \le 0] = 0$ and for $x > 0$ we have

$$\mathbb{P}[X \le x] = \int_{-\infty}^x f_X(u)\,du = \int_0^x \lambda e^{-\lambda u}\,du = 1 - e^{-\lambda x}.$$

Therefore,

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

A sketch of the distribution function $F_X(x)$ looks like:



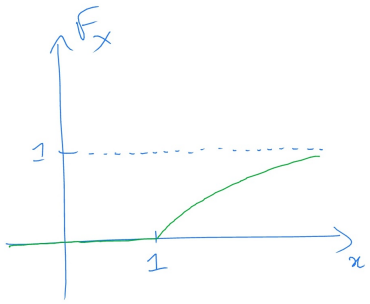
Example 4: *Properties of distribution functions*

> *Let*

$$F(x) = \begin{cases} 1 - \frac{1}{x} & \text{if } x > 1 \\ 0 & \text{otherwise.} \end{cases}$$

Sketch F and show that F is a distribution function

A sketch of F looks like



To show that F is a distribution function, we'll check properties 1-3 from Section 1.2.

1. From the definition, $0 \le F(x) \le 1$ for all x. Since $F(x) = 0$ for all $x \le 1$ we have $\lim_{x\to-\infty} F(x) = 0$, and also $\lim_{x\to\infty} 1 - \frac{1}{x} = 1 - 0 = 1$.

2. Since $F(x) = 0$ for all $x \le 1$, its clear that $F(x)$ is non-decreasing while $x \le 1$. If $0 \le x < y$ then $\frac{1}{y} < \frac{1}{x}$ so $1 - \frac{1}{x} \le 1 - \frac{1}{y}$. Hence F is non-decreasing across all $x \in \mathbb{R}$.

3. From its definition, F is continuous for $x \in (-\infty, 1)$ and $x \in (1, \infty)$. Since $F(1+) = 1 - \frac{1}{1} = 0 = F(1) = F(1-)$, we have that F is continuous everywhere. (Alternatively, in this course, we allow ourselves to 'prove' continuity by drawing a sketch, as above.)

Hence, F is a distribution function, and as a result there exists a random variable X with distribution function $F_X = F$.

Example 5: *Calculating expectations and variances*

> *Let X be an Exponential random variable, from Example 3, with p.d.f.*

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Find the mean and variance of X.

We can calculate, integrating by parts,

$$\begin{aligned} \mathbb{E}[X] &= \int_{-\infty}^{\infty} x f_X(x)\, dx = \int_0^{\infty} x\lambda e^{-\lambda x} = \Big[x(-1)e^{-\lambda x}\Big]_{x=0}^{\infty} - \int_0^{\infty} -e^{-\lambda x}\, dx \\ &= 0 + \left[\frac{-1}{\lambda} e^{-\lambda x}\right]_{x=0}^{\infty} = \frac{1}{\lambda}. \end{aligned}$$

For the variance, it is easiest to calculate $\mathbb{E}[X^2]$ and then use (from MAS113) that $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$. So,

$$\begin{aligned} \mathbb{E}[X^2] &= \int_{-\infty}^{\infty} x^2 f_X(x)\, dx = \int_0^{\infty} x^2\lambda e^{-\lambda x} = \Big[x^2(-1)e^{-\lambda x}\Big]_{x=0}^{\infty} - \int_0^{\infty} 2x(-1)e^{-\lambda x}\, dx \\ &= 0 + \frac{2}{\lambda}\int_0^{\infty} x\lambda e^{-\lambda x}\, dx = \frac{2}{\lambda^2} \end{aligned}$$

where we use that we already calculated $\int_0^{\infty} x\lambda e^{-\lambda x}\, dx = \frac{1}{\lambda}$. Hence,

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{1}{\lambda^2}.$$

Chapter 2

Example 6: *Calculating $\mathbb{E}[e^Y]$ where $X \sim N(0,1)$.*

> *Let Y be a normal random variable, with mean 0 and variance 1, with p.d.f.*

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}.$$

Find $\mathbb{E}[e^Y]$.

We need to calculate

$$\mathbb{E}[e^Y] = \int_{-\infty}^{\infty} e^y f_Y(y)\, dy = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^y e^{-y^2/2}\, dy. \tag{1}$$

We can't evaluate this integral explicitly. However, we do know the value of a similar integral, that is we know

$$\mathbb{P}[Y \in \mathbb{R}] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{y^2/2}\, dy = 1. \tag{2}$$

Our aim is to rewrite $\mathbb{E}[e^Y]$ into this form (and hope we can deal with whatever else is left over). We can do so by completing the square:

$$e^y e^{-y^2/2} = \exp\left(-\frac{y^2}{2} + y - \frac{1}{2} + \frac{1}{2}\right) = \exp\left(-\frac{1}{2}(y-1)^2 + \frac{1}{2}\right) = e^{-(y-1)^2/2} e^{1/2}.$$

Putting this into (1), we have

$$\begin{aligned}\mathbb{E}[e^Y] &= e^{1/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(y-1)^2/2}\, dy \\ &= e^{1/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz\end{aligned}$$

where $z = y - 1$. Then, using (2) we have $\mathbb{E}[e^Y] = e^{1/2}$.

See **Q2.9** for a more general case of this method.

Example 7: *Mean and variance of the Gamma distribution*

> *Let X have the $Ga(\alpha, \beta)$ distribution, where $\alpha, \beta > 0$. Find the mean and variance of X.*

We can calculate

$$\begin{aligned}\mathbb{E}[X] &= \int_{-\infty}^{\infty} x f_X(x)\, dx \\ &= \int_0^{\infty} \frac{\beta^\alpha}{\Gamma(\alpha)} x^\alpha e^{-\beta x}\, dx \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+1)}{\beta^{\alpha+1}} \qquad \text{using Lemma 2.3} \\ &= \frac{\alpha\Gamma(\alpha)}{\beta\Gamma(\alpha)} \qquad \text{using Lemma 2.2} \\ &= \frac{\alpha}{\beta}.\end{aligned}$$

Similarly, for the variance,

$$\begin{aligned}\mathbb{E}[X^2] &= \int_{-\infty}^{\infty} x^2 f_X(x)\, dx \\ &= \int_0^{\infty} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha+1} e^{-\beta x}\, dx \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+2)}{\beta^{\alpha+2}} \qquad \text{using Lemma 2.3} \\ &= \frac{\alpha(\alpha+1)\Gamma(\alpha)}{\beta^2\Gamma(\alpha)} \qquad \text{using Lemma 2.2} \\ &= \frac{\alpha(\alpha+1)}{\beta^2}.\end{aligned}$$

So

$$\mathrm{Var}(X) = \mathbb{E}[X^2] = \mathbb{E}[X]^2 = \frac{\alpha(\alpha+1) - \alpha^2}{\beta^2} = \frac{\alpha}{\beta^2}.$$

Example 8: *Mean and variance of the Beta distribution*

> *Let X have the $Be(\alpha, \beta)$ distribution, where $\alpha, \beta > 0$. Find the mean and variance of X.*

For the mean,

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^1 \frac{1}{B(\alpha,\beta)} x^{\alpha}(1-x)^{\beta-1}\, dx \\
&= \frac{B(\alpha+1,\beta)}{B(\alpha,\beta)} \\
&= \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} \Big/ \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \qquad \text{using (2.5)} \\
&= \frac{\alpha\Gamma(\alpha)\Gamma(\alpha+\beta)}{(\alpha+\beta)\Gamma(\alpha+\beta)\Gamma(\alpha)} \qquad \text{using Lemma 2.2} \\
&= \frac{\alpha}{\alpha+\beta}.
\end{aligned}$$

For the variance,

$$\begin{aligned}
\mathbb{E}[X^2] &= \int_0^1 \frac{1}{B(\alpha,\beta)} x^{\alpha+1}(1-x)^{\beta-1}\, dx \\
&= \frac{B(\alpha+2,\beta)}{B(\alpha,\beta)} \\
&= \frac{\Gamma(\alpha+2)\Gamma(\beta)}{\Gamma(\alpha+\beta+2)} \Big/ \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \qquad \text{using (2.5)} \\
&= \frac{\alpha(\alpha+1)\Gamma(\alpha)\Gamma(\alpha+\beta)}{(\alpha+\beta)(\alpha+\beta+1)\Gamma(\alpha+\beta)\Gamma(\alpha)} \qquad \text{using Lemma 2.2} \\
&= \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}.
\end{aligned}$$

So, using that $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$ we have

$$\begin{aligned}
\mathrm{Var}(X) &= \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} - \frac{\alpha^2}{(\alpha+\beta)^2} \\
&= \frac{\alpha(\alpha+1)(\alpha+\beta) - \alpha^2(\alpha+\beta+1)}{(\alpha+\beta)^2(\alpha+\beta+1)} \\
&= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.
\end{aligned}$$

Chapter 3

Example 9: *Cube root of the $Be(3,1)$ distribution.*

> *Let $X \sim Be(3,1)$ and let $Y = \sqrt[3]{X}$. Find the probability density function of Y.*

From (2.6), the p.d.f. of the $Be(\alpha,\beta)$ distribution is

$$f_X(x) = \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} & \text{if } x \in (0,1) \\ 0 & \text{otherwise.} \end{cases}$$

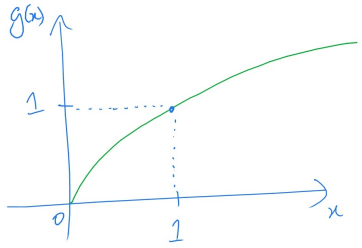
Note that, for any $\alpha > 1$,

$$B(\alpha, 1) = \frac{\Gamma(\alpha)\Gamma(1)}{\Gamma(\alpha+1)} = \frac{\Gamma(\alpha)\Gamma(1)}{\alpha\Gamma(\alpha)} = \frac{1}{\alpha} \tag{3}$$

by (2.5) and Lemma 2.2. Putting this, along with $\alpha = 3$ and $\beta = 1$ into (2.6), the p.d.f. of X is

$$f_X(x) = \begin{cases} 3x^2 & \text{if } x \in (0,1) \\ 0 & \text{otherwise.} \end{cases}$$

For the transformation, we use the function $g(x) = \sqrt[3]{x}$, which is strictly increasing.



The p.d.f. of X is non-zero on $(0,1)$, and g maps $R_X = (0,1)$ to $(0,1)$, so $g(R_X) = (0,1)$. We have $g^{-1}(y) = y^3$, so $\frac{dg^{-1}}{dy} = 3y^2$. Therefore, by Lemma 3.1 we have

$$f_Y(y) = \begin{cases} 3(y^3)^2 \times 3y^2 & \text{if } y \in (0,1) \\ 0 & \text{otherwise,} \end{cases}$$

$$= \begin{cases} 9y^8 & \text{if } y \in (0,1) \\ 0 & \text{otherwise.} \end{cases}$$

In fact, using the same calculations as above, it can be seen that this is the p.d.f. of a $Be(9,1)$ distribution. See **Q3.5** for a more general case.

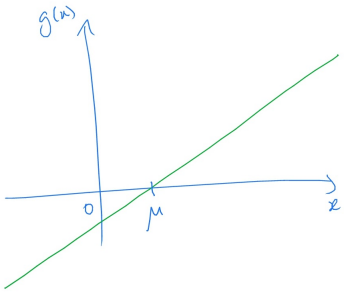
Example 10: *Standardization of the normal distribution.*

> *Let $X \sim N(\mu, \sigma^2)$ and define $Y = \frac{X-\mu}{\sigma}$. Show that $Y \sim N(0,1)$.*

The p.d.f. of the normal distribution with mean μ and variance σ^2 is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

with range $R_X = \mathbb{R}$. The function $g(x) = \frac{x-\mu}{\sigma}$ is strictly increasing, and $g(\mathbb{R}) = \mathbb{R}$.



If $y = \frac{x-\mu}{\sigma}$ then $x = \sigma y + \mu$, hence the inverse function is $g^{-1}(y) = \sigma y + \mu$, with derivative $\frac{dg^{-1}}{dy} = \sigma > 0$. Hence, by Lemma 3.1,

$$\begin{aligned} f_Y(y) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{y^2}{2}\right) \times \sigma \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) \end{aligned}$$

which is the p.d.f. of a $N(0,1)$ random variable.

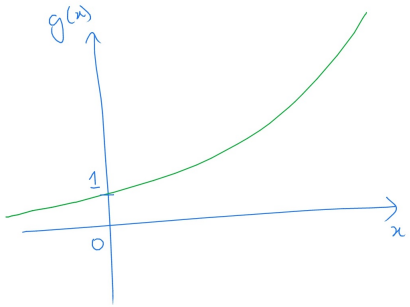
Example 11: *The log-normal distribution.*

> *Find the probability density function of $Y = e^X$, where $X \sim N(\mu, \sigma^2)$.*

Recall that Y is known as the log-normal distribution, which we introduced in Section 2.2.2. The probability density function of X is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

which is non-zero for all $x \in \mathbb{R}$. Our transformation is $g(x) = e^x$, which is strictly increasing for all $x \in \mathbb{R}$.



The range of X is $\mathbb{R}$, which is mapped by g to $g(\mathbb{R}) = (0, \infty)$.

We have $g^{-1}(y) = \log y$, and $\frac{dg^{-1}}{dy} = \frac{1}{y}$. Hence, by Lemma 3.1 the p.d.f. of Y is given by

$$f_Y(y) = \begin{cases} \frac{1}{y\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log y - \mu)^2}{2\sigma^2}\right) & \text{if } y \in (0, \infty) \\ 0 & \text{otherwise.} \end{cases}$$

Example 12: *Square of a standard normal (the chi-squared distribution).*

> *Let $X \sim N(0,1)$ and let $Y = X^2$. Find the p.d.f. of Y and verify that Y has the χ^2_1 distribution.*

We aim to find the p.d.f. of Y and check that it matches the p.d.f. given for the χ^2_1 distribution in Section 2.3.3. Note that $R_X = \mathbb{R}$, and we can't apply Lemma 3.1 because $g(x) = x^2$ is not strictly monotone on $\mathbb{R}$.

If $y < 0$ then $\mathbb{P}[Y \leq y] = 0$ because $Y = X^2 \geq 0$. Moreover, because the normal distribution is a continuous distribution, $\mathbb{P}[X = 0] = 0$, so also $\mathbb{P}[Y \leq 0] = 0$

This leaves $y > 0$, and in this case we have

$$\begin{aligned} F_Y(y) = \mathbb{P}[Y \leq y] &= \mathbb{P}[-\sqrt{y} \leq X \leq \sqrt{y}] \\ &= \mathbb{P}[X \leq \sqrt{y}] - \mathbb{P}[X \leq -\sqrt{y}] \\ &= \Phi(\sqrt{y}) - \Phi(-\sqrt{y}) \end{aligned}$$

Here, $\Phi(x) = \mathbb{P}[X \leq x]$ is the distribution function of the standard normal distribution. Differentiating with respect to y, we have

$$\begin{aligned} f_Y(y) &= \frac{1}{2\sqrt{y}}\phi(\sqrt{y}) - \frac{-1}{2\sqrt{y}}\phi(-\sqrt{y}) \\ &= \frac{1}{\sqrt{y}}\phi(\sqrt{y}) \\ &= \frac{1}{\sqrt{2\pi y}}\exp(-y/2). \end{aligned}$$

Here, ϕ is the probability density function of the standard normal distribution. We use that $\phi(x) = \phi(-x)$.

If we recall (from Section 2.3.1) that $\Gamma(1/2) = \sqrt{\pi}$, we then have

$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{2}\Gamma(1/2)}\frac{1}{\sqrt{y}}\exp\left(-\frac{y}{2}\right) & \text{if } y > 0 \\ 0 & \text{otherwise.} \end{cases}$$

which exactly matches the p.d.f. given for the χ^2_1 distribution in Section 2.3.3.

Chapter 4

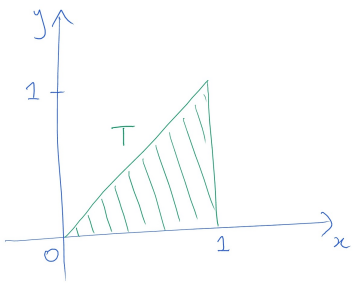
Example 13: *Joint probability density functions*

> *Let T be the triangle $\{(x, y) : x \in (0, 1), y \in (0, x)\}$. Define*

$$f(x, y) = \begin{cases} k(x + y) & \text{if } (x, y) \in T \\ 0 & \text{otherwise.} \end{cases}$$

Find the value of k such that f is a joint probability density function.

First, we sketch the region T on which $f_{X,Y}(x, y)$ is non-zero.



We need $f(x,y) \geq 0$ for all x, y, which means we must have $k \geq 0$. Also, we need that $\iint_T f(x,y)\,dx\,dy = 1$. Therefore,

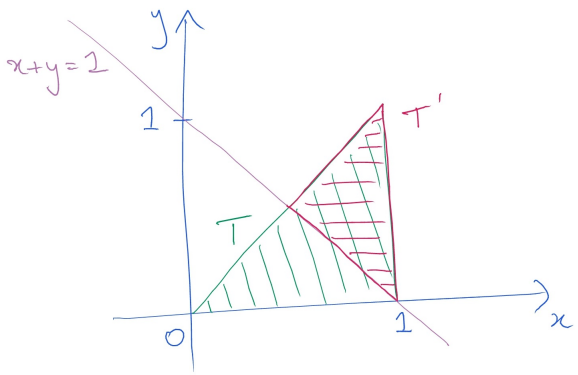
$$\begin{aligned} 1 &= \iint_T f(x,y)\,dy\,dx \\ &= \int_0^1 \int_0^x k(x+y)\,dy\,dx \\ &= k\int_0^1 \left[xy + \frac{y^2}{2}\right]_{y=0}^{x} dx \\ &= k\int_0^1 \frac{3x^2}{2}\,dx \\ &= \frac{k}{2}. \end{aligned}$$

So $k = 2$.

Here, to find the limits of integration, we describe the region T as being covered by vertical lines, one for each fixed x. With x fixed, the range of y that makes up T is $y \in (0, x)$. That is, we use that $T = \{(x,y) \ : \ x \in (0,1), y \in (0,x)\}$.

> *If X and Y have joint p.d.f. $f_{X,Y}(x,y) = f(x,y)$, find $\mathbb{P}[X+Y>1]$.*

To find $\mathbb{P}[X+Y>1]$, we need to integrate $f_{X,Y}(x,y)$ over the region of (x,y) for which $(x,y) \in T$ and $x+y>1$. Let's call this region T', and sketch it.



We have $T' = \{(x,y) \ : \ x \in (\frac{1}{2},1), y \in (1-x,x)\}$. So,

$$\begin{aligned}
\mathbb{P}[X+Y>1] &= \int_{\frac{1}{2}}^{1} \int_{1-x}^{x} 2(x+y)\,dy\,dx \\
&= \int_{\frac{1}{2}}^{1} \left[2xy+y^2\right]_{y=1-x}^{x} dx \\
&= \int_{\frac{1}{2}}^{1} (4x^2-1)\,dx \\
&= \left[\frac{4}{3}x^3 - x\right]_{\frac{1}{2}}^{1} \\
&= \frac{1}{3} - \left(-\frac{1}{3}\right) \\
&= \frac{2}{3}
\end{aligned}$$

Example 14: *Marginal distributions*

> *Let (X,Y) be as in Example 13. Find the marginal p.d.f.s of X and Y.*

For $x \in (0,1)$,

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy = \int_0^x 2(x+y)\,dy = \left[2xy+y^2\right]_{y=0}^{x} = 3x^2.$$

Here, to find the limits of the integral, we keep x fixed, and then look for the range of y for which $f_{X,Y}(x,y)$ is non-zero. That is, we use $T = \{(x,y) \ : \ x \in (0,1), y \in (0,x)\}$.

For $x \notin (0,1)$, we have $f_{X,Y}(x,y) = 0$, so

$$f_X(x) = \begin{cases} 3x^2 & \text{if } x \in (0,1) \\ 0 & \text{otherwise.} \end{cases}$$

is the marginal p.d.f. of X.

For $y \in (0,1)$, we have

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx = \int_y^1 2(x+y)\,dx = \left[x^2+2xy\right]_{x=y}^{1} = 1+2y-3y^2.$$

Here, to find the limits of the integral, we keep y fixed, and then look for the range of x for which $f_{X,Y}(x,y)$ is non-zero. That is, we use $T = \{(x,y) \ : \ y \in (0,1), x \in (y,1)\}$.

For $y \notin (0,1)$ we have $f_{X,Y}(x,y) = 0$, so

$$f_Y(y) = \begin{cases} 1+2y-3y^2 & \text{if } y \in (0,1) \\ 0 & \text{otherwise.} \end{cases}$$

is the marginal p.d.f. of Y.

Example 15: *Conditional distributions*

> *Let (X, Y) be as in Example 13. For $y \in (0, 1)$, find the conditional p.d.f. of X given $Y = y$.*

We obtained $f_Y(y)$ in Example 14, and we know $f_{X,Y}(x, y)$ from Example 13. Note that, with $y \in (0, 1)$ fixed, $f_{X,Y}(x, y)$ is non-zero only for $x \in (y, 1)$. So,

$$f_{X|Y=y}(x) = \frac{f_{X,Y}(x, y)}{f_Y(y)} = \begin{cases} \frac{2(x+y)}{1+2y-3y^2} & \text{if } x \in (y, 1) \\ 0 & \text{otherwise.} \end{cases}$$

Example 16: *Independence, factorizing $f_{X,Y}$.*

> *Are the random variables X and Y from Example 16 independent?*

The random variables X and Y from Example 13 are not independent as the p.d.f.

$$f(x, y) = \begin{cases} 2(x + y) & \text{if } (x, y) \in T \\ 0 & \text{otherwise} \end{cases}$$

cannot be factorised as a function of x times a function of y.

> *Let U and V be two random variables with joint probability density function*

$$f_{U,V}(u, v) = \begin{cases} 12ue^{-(2u+3v)} & \text{if } u > 0, v > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Are U and V independent?

$f_{U,V}(u, v)$ can be factorised into a function of x and a function of y,

$$\begin{aligned} f_{U,V}(u, v) &= \begin{cases} 4ue^{-2u} \cdot 3e^{-3v} & \text{if } u > 0, v > 0 \\ 0 & \text{otherwise.} \end{cases} \\ &= g(u)h(v) \end{aligned}$$

where

$$g(u) = \begin{cases} 4ue^{-2u} & \text{if } u > 0 \\ 0 & \text{otherwise,} \end{cases} \qquad h(v) = \begin{cases} 3e^{-3v} & \text{if } v > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, U and V are independent.

In fact, in this case we can recognize that g is the p.d.f. of a $Ga(2, 2)$ and h is the p.d.f. of a $Exp(3)$, so U and Y are $Ga(2, 2)$ and $Exp(3)$ respectively.

Example 17: *Covariance and correlation*

> *Let (X, Y) be as in Example 13. Find the covariance $\mathrm{Cov}(X, Y)$.*

We start by calculating $\text{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$. We have

$$\begin{aligned}\mathbb{E}[XY] &= \int_0^1 \int_0^x 2xy(x+y)\,dy\,dx \\ &= \int_0^1 \frac{5}{3}x^4\,dx \\ &= \frac{1}{3}.\end{aligned}$$

Using the marginal probability density functions for X and Y that we found in Example 14, we have

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\,dx = \int_0^1 x(3x^2)\,dx = \frac{3}{4}$$
$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} y f_Y(y)\,dy = \int_0^1 y(1+2y-3y^2)\,dy = \frac{5}{12}$$

(evaluating these two integrals is left to you). So,

$$\text{Cov}(X,Y) = \frac{1}{3} - \frac{3}{4}\cdot\frac{5}{12} = \frac{1}{48}.$$

> *Find the correlation $\rho(X,Y)$.*

We now need to find $\rho(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}}$. So, we also need to calculate the variances of X and Y. We have

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x)\,dx = \int_0^1 x^2(3x^2)\,dx = \frac{3}{5}$$
$$\mathbb{E}[Y^2] = \int_{-\infty}^{\infty} y^2 f_Y(y)\,dy = \int_0^1 y^2(1+2y-3y^2)\,dy = \frac{7}{30}$$

(again, evaluating these two integrals is left to you). From this we obtain,

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{3}{5} - \left(\frac{3}{4}\right)^2 = \frac{3}{80}$$
$$\text{Var}(Y) = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = \frac{7}{30} - \left(\frac{5}{12}\right)^2 = \frac{43}{720}$$

and we get

$$\rho(X,Y) = \frac{1/48}{\sqrt{\frac{3}{80}\cdot\frac{43}{720}}} \approx 0.44$$

Example 18: *Calculating conditional expectation*

> *Let (X,Y) be as in Example 13. Let $y \in (0,1)$. Find $\mathbb{E}[X|Y=y]$ and $\mathbb{E}[X|Y]$.*

We have already found the conditional p.d.f. of X in Example 15, it is

$$f_{X|Y=y}(x) = \begin{cases} \frac{2(x+y)}{1+2y-3y^2} & \text{if } x \in (y,1) \\ 0 & \text{otherwise.} \end{cases}$$

So,

$$\mathbb{E}[X|Y=y] = \int_y^1 \frac{2(x^2+yx)}{1+2y-3y^2}\,dx = \left[\frac{\frac{2}{3}x^3+yx^2}{1+2y-3y^2}\right]_y^1 = \frac{2+3y-5y^3}{3(1+2y-3y^2)}.$$

Hence,

$$\mathbb{E}[X|Y] = \frac{2+3Y-5Y^3}{3(1+2Y-3Y^2)}.$$

> *Show that* $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$.

To find $\mathbb{E}[\mathbb{E}[X|Y]]$, we first note that

$$\mathbb{E}[X|Y] = g(Y) = \frac{2+3Y-5Y^3}{3(1+2Y-3Y^2)}$$

use then use the usual method for finding the expectation of a function of Y. That is,

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[g(Y)] = \int_0^1 g(y)f_Y(y)\,dy = \int_0^1 \frac{2+3y-5y^3}{3}\,dy = \frac{2}{3}+\frac{1}{2}-\frac{5}{12} = \frac{3}{4}.$$

We have already shown during Example 17 that $\mathbb{E}[X] = \frac{3}{4}$.

Example 19: *Proof of* $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$

It is no coincidence that $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$ in Example 18. In fact, this holds true for all pairs of random variables X and Y. Here is a general proof.

We have $\mathbb{E}[X|Y] = g(Y)$, where

$$g(y) = \mathbb{E}[X|Y=y] = \int_{-\infty}^{\infty} x f_{X|Y=y}(x)\,dx.$$

So,

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[X|Y]] &= \mathbb{E}[g(Y)] \\
&= \int_{-\infty}^{\infty} g(y)f_Y(y)\,dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X|Y=y}(x)f_Y(y)\,dx\,dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X,Y}(x,y)\,dy\,dx \qquad \text{(by definition of the conditional p.d.f.)} \\
&= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy\,dx \\
&= \int_{-\infty}^{\infty} x f_X(x)\,dx \qquad \text{(by definition of the marginal p.d.f.)} \\
&= \mathbb{E}[X].
\end{aligned}$$

Example 20: *Calculation of expectation and variance by conditioning*

Let $X \sim Ga(2,2)$ and, conditional on $X = x$, let $Y \sim Po(x)$. Then, using standard results about the mean and variance of Gamma/Poisson random variables, $\mathbb{E}[X] = 1, \text{Var}(X) = \frac{1}{2}, \mathbb{E}[Y|X] = X$ and $\text{Var}(Y|X) = X$. So, using the formulae from Lemma 4.10,

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[X] = 1$$
$$\text{Var}(Y) = \mathbb{E}[\text{Var}(Y|X)] + \text{Var}(\mathbb{E}[Y|X]) = \mathbb{E}[X] + \text{Var}(X) = \frac{3}{2}.$$

Chapter 5

Example 21: *Transforming bivariate random variables*

> *Let $X \sim Ga(3,1)$ and $Y \sim Be(2,2)$, and let X and Y be independent. Find the joint p.d.f. of the vector (U,V), where $U = X + Y$ and $V = X - Y$.*

The p.d.f.s of X and Y are

$$f_X(x) = \begin{cases} \frac{1}{2}x^2e^{-x} & \text{if } x > 0 \\ 0 & \text{otherwise,} \end{cases} \qquad f_Y(y) = \begin{cases} 6y(1-y) & \text{if } y \in (0,1) \\ 0 & \text{otherwise.} \end{cases}$$

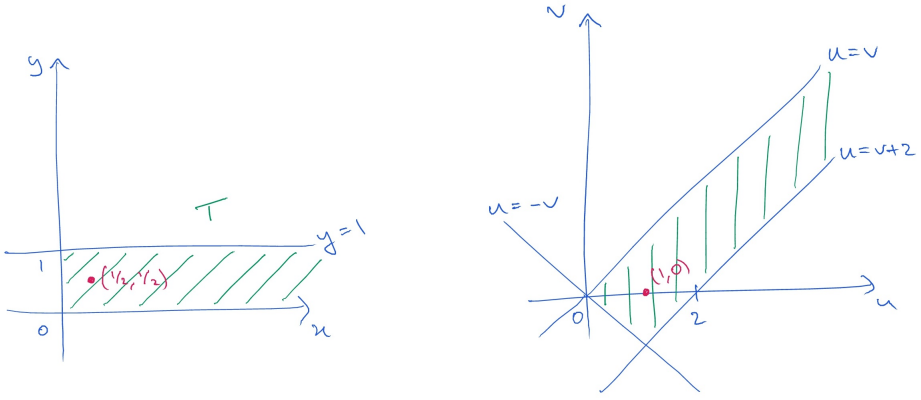
By independence, their joint p.d.f. is

$$f_{X,Y}(x,y) = \begin{cases} 3x^2y(1-y)e^{-x} & \text{if } x > 0 \text{ and } y \in (0,1) \\ 0 & \text{otherwise.} \end{cases}$$

The transformation we want is $u = x + y$ and $v = x - y$. So, $u + v = 2x$, $u - v = 2y$, and the inverse transformation is $x = \frac{u+v}{2}$, and $y = \frac{u-v}{2}$. Hence, the Jacobian is

$$J = \det \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{pmatrix} = -\frac{1}{2}.$$

Now, we need to transform the region $T = \{(x,y) \ : \ x > 0, y \in (0,1)\}$ into the (u,v) plane. This region is bounded by the three lines $x = 0$, $y = 0$ and $y = 1$, which map respectively to the lines $u = -v$, $u = v$ and $u = v + 2$.



Our transformed region must also be bounded by the three lines; to check which section of the sketch it is we simply find out where some $(x, y) \in T$ maps to. We have $(\frac{1}{2}, \frac{1}{2}) \in T$ which maps to $(1, 0)$, so the shaded region is the image of T.

Therefore,

$$f_{U,V}(u,v) = \begin{cases} f_{X,Y}\left(\frac{u+v}{2}, \frac{u-v}{2}\right) \times \left|\frac{-1}{2}\right| & \text{if } u > 0, v \in (u-2, u), v > -u \\ 0 & \text{otherwise.} \end{cases}$$

$$= \begin{cases} \frac{3}{32}(u+v)^2(u-v)(2-u+v)e^{\frac{u+v}{2}} & \text{if } u > 0, v \in (u-2, u), v > -u \\ 0 & \text{otherwise.} \end{cases}$$

Example 22: *The Box-Muller transform, simulation of normal random variables*

Let $S \sim Exp(\frac{1}{2})$ and $\Theta \sim U[0, 2\pi)$, and let S and Θ be independent. Then S and Θ have joint p.d.f. given by

$$f_{S,\Theta}(s,\theta) = \begin{cases} \frac{1}{4\pi}e^{-\frac{1}{2}s} & \text{if } s \geq 0 \text{ and } \theta \in [0, 2\pi) \\ 0 & \text{otherwise.} \end{cases}$$

We can think of S and Θ as giving the location of a point $(\sqrt{S}, \Theta)$ in polar co-ordinates. We transform this point into Cartesian co-ordinates, meaning that we want to use the transformation $X = \sqrt{S}\cos(\Theta)$ and $Y = \sqrt{S}\sin(\Theta)$. Therefore, our transformation is

$$x = \sqrt{s}\cos\theta, \qquad y = \sqrt{s}\sin\theta.$$

This transformation maps the set of (s, θ) for which $f_{S,\Theta}(s, \theta) > 0$ onto all of $\mathbb{R}^2$ (it is just Polar coordinates (r, θ) with $r = \sqrt{s}$).

To find the inverse transformation, note that $s = x^2 + y^2$ and $y/x = \tan\theta$, so $\theta = \arctan(y/x)$. So the Jacobian is

$$J = \det\begin{pmatrix} \frac{\partial s}{\partial x} & \frac{\partial s}{\partial y} \\ \frac{\partial \theta}{\partial x} & \frac{\partial \theta}{\partial y} \end{pmatrix} = \det\begin{pmatrix} 2x & 2y \\ \frac{-y/x^2}{1+(y/x)^2} & \frac{1/x}{1+(y/x)^2} \end{pmatrix} = \frac{2}{1+(y/x)^2} - \frac{-2y^2/x^2}{1+(y/x)^2} = 2$$

Hence,

$$f_{X,Y}(x,y) = \frac{1}{2\pi}e^{-\frac{x^2+y^2}{2}}$$

for all $(x, y) \in \mathbb{R}^2$. Now, we can factorise this as

$$f_{X,Y}(x,y) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}},$$

which implies that X and Y are independent standard normal random variables.

Assuming we can simulate uniform random variables, then using the transformation in **Q3.3** we can also simulate exponential random variables. Then, using above transformation, we can simulate standard normals.

Example 23: *Finding the distribution of a sum of Gamma random variables*

> *Suppose that two independent random variables X and Y follow the distributions $X \sim Ga(4,2)$ and $Y \sim Ga(2,2)$. Find the distribution of $Z = X + Y$*

Let $W = X$. So the transformation we want to apply is

$$z = x + y, \qquad w = x.$$

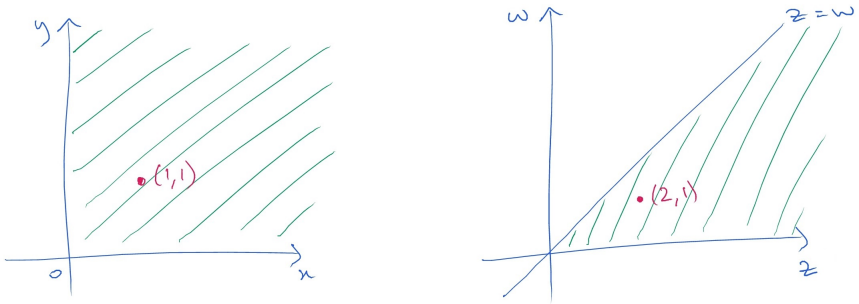
The inverse transformation is $x = w$ and $y = z - w$, so the Jacobian is

$$J = \det \begin{pmatrix} \frac{\partial x}{\partial z} & \frac{\partial x}{\partial w} \\ \frac{\partial y}{\partial z} & \frac{\partial y}{\partial w} \end{pmatrix} = \det \begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix} = -1.$$

By independence of X and Y, their joint p.d.f. is

$$f_{X,Y}(x,y) = \begin{cases} \frac{2^4}{\Gamma(4)} x^3 e^{-2x} \frac{2^2}{\Gamma(2)} y e^{-2y} & \text{if } x, y > 0 \\ 0 & \text{otherwise} \end{cases}$$
$$= \begin{cases} \frac{2^6}{6} x^3 y e^{-2(x+y)} & \text{if } x, y > 0 \\ 0 & \text{otherwise.} \end{cases}$$

The region of (x, y) on which $f_{X,Y}(x,y)$ is non-zero is $x > 0$ and $y > 0$. This is bounded by the lines $x = 0$, $y = 0$, which are respectively mapped to $w = 0$ and $z = w$.



The point $(1, 1)$ is mapped to $(2, 1)$, meaning that the shaded area is the region on which $f_{Z,W}(z,w)$ is non-zero.

Hence, the joint p.d.f. of Z and W is

$$f_{Z,W}(z,w) = \begin{cases} \frac{2^6}{6} w^3 (z-w) e^{-2z} & \text{if } z > 0 \text{ and } w \in (0, z) \\ 0 & \text{othwerwise.} \end{cases}$$

Lastly, to obtain the marginal p.d.f. of Z, we integrate out w. For $z > 0$,

$$f_Z(z) = \frac{2^6}{6} e^{-2z} \int_0^z (w^3 z - w^4)\, dw$$
$$= \frac{2^6}{6} e^{-2z} \left(\frac{z^5}{4} - \frac{z^5}{5} \right)$$
$$= \frac{2^6}{6 \cdot 20} z^5 e^{-2z}$$
$$= \frac{2^6}{\Gamma(6)} z^5 e^{-2z}.$$

For $z \leq 0$ we have $f_Z(z) = 0$. So, we can recognise $f_Z(z)$ as the p.d.f. of a $Ga(6,2)$ random variable, and conclude that $Z \sim Ga(6,2)$.

More generally, this method can be used to show that if $X \sim Ga(\alpha_1, \beta), Y \sim Ga(\alpha_2, \beta)$ and X and Y are independent, then $X + Y \sim Ga(\alpha_1 + \alpha_2, \beta)$ for any $\alpha_1, \alpha_2, \beta$. See **Q5.8**.

Chapter 6

Example 24: *Mean vectors and covariance matrices*

Recall the random variables (X, Y) from Example 13. In Example 17 we calculated that $\mathbb{E}[X] = \frac{3}{4}$ and $\mathbb{E}[Y] = \frac{5}{12}$. So the mean vector of $\mathbf{X} = (X, Y)^T$ is

$$\mathbb{E}[\mathbf{X}] = \begin{pmatrix} \frac{3}{4} \\ \frac{5}{12} \end{pmatrix}.$$

In Example 17 we also calculated that $\text{Cov}(X, Y) = \frac{1}{48}$ and that $\text{Var}[X] = \frac{3}{80}$, $\text{Var}(Y) = \frac{43}{720}$. Therefore, the covariance matrix of $\mathbf{X}$ is

$$\text{Cov}(\mathbf{X}) = \begin{pmatrix} \frac{3}{80} & \frac{1}{48} \\ \frac{1}{48} & \frac{43}{720} \end{pmatrix}.$$

Example 25: *Affine transformation of a random vector*

> *Suppose that the random vector* $\mathbf{X} = (X_1, X_2, X_3)^T$ *has*
>
> $$\mathbb{E}[\mathbf{X}] = \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}, \qquad \text{Cov}(\mathbf{X}) = \begin{pmatrix} 2 & 0 & 1 \\ 0 & 2 & \frac{2}{3} \\ 1 & \frac{2}{3} & 2 \end{pmatrix}.$$
>
> *Define two new random variables,* $U = X_1 - X_2 + X_3$ *and* $V = X_1 - X_3 + 1$. *Find the mean vector and covariance matrix of* $\mathbf{U} = (U, V)^T$.

We can express the relationship between $\mathbf{X}$ and $\mathbf{U}$ as an affine transformation:

$$\mathbf{U} = \begin{pmatrix} U \\ V \end{pmatrix} = \mathbf{A}\mathbf{X} + \mathbf{b} = \begin{pmatrix} 1 & -1 & 1 \\ 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

So, we can use Lemma 6.3 to find the mean vector and covariance matrix of $\mathbf{U}$. Firstly,

$$\begin{aligned} \mathbb{E}[\mathbf{U}] &= \mathbf{A}\mathbb{E}[\mathbf{X}] + \mathbf{b} \\ &= \begin{pmatrix} 1 & -1 & 1 \\ 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} 2 \\ -1 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} 2 \\ 0 \end{pmatrix} \end{aligned}$$

and secondly,

$$\begin{aligned}\mathrm{Cov}(\mathbf{U}) &= \mathbf{A}\,\mathrm{Cov}(\mathbf{X})\mathbf{A}^T \\ &= \begin{pmatrix} 1 & -1 & 1 \\ 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} 2 & 0 & 1 \\ 0 & 2 & \frac{2}{3} \\ 1 & \frac{2}{3} & 2 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ -1 & 0 \\ 1 & -1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & -1 & 1 \\ 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} 3 & 1 \\ -\frac{4}{3} & -\frac{2}{3} \\ \frac{7}{3} & -1 \end{pmatrix} \\ &= \begin{pmatrix} \frac{20}{3} & \frac{2}{3} \\ \frac{2}{3} & 2 \end{pmatrix}.\end{aligned}$$

> *Find the correlation coefficient $\rho(U, V)$.*

We can read off $\mathrm{Var}(U), \mathrm{Var}(V)$ and $\mathrm{Cov}(U, V)$ from the covariance matrix of $\mathbf{U}$. So the correlation coefficient of U and V is

$$\rho(U, V) = \frac{\mathrm{Cov}(U, V)}{\sqrt{\mathrm{Var}(U)\,\mathrm{Var}(V)}} = \frac{\frac{2}{3}}{\sqrt{\frac{20}{3} \cdot 2}} = \frac{1}{\sqrt{30}}.$$

Example 26: *Variance of a sum*

> *Suppose that two random variables X and Y have variances σ_X^2 and σ_Y^2, and covariance $\mathrm{Cov}(X, Y)$. Find the variance of $X + Y$.*

If we write $U = X + Y$, then

$$\begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = (U) = U$$

where (U) denotes the 1×1 matrix with the single entry U. We usually won't bother to write brackets around 1×1 matrices/vectors. We can apply Lemma 6.3 to this case, with $\mathbf{A} = \begin{pmatrix} 1 & 1 \end{pmatrix}$ and $\mathbf{X} = (X, Y)^T$, to obtain that

$$\mathrm{Cov}(U) = \mathbf{A}\,\mathrm{Cov}(\mathbf{X})\mathbf{A}^T.$$

The covariance matrix of $\mathbf{X}$ is given by

$$\mathrm{Cov}(\mathbf{X}) = \begin{pmatrix} \sigma_X^2 & \mathrm{Cov}(X, Y) \\ \mathrm{Cov}(X, Y) & \sigma_Y^2 \end{pmatrix}.$$

Since U is 1×1, $\mathrm{Cov}(U) = \mathrm{Var}(U)$, so we have

$$\begin{aligned}\mathrm{Var}(X + Y) &= \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} \sigma_X^2 & \mathrm{Cov}(X, Y) \\ \mathrm{Cov}(X, Y) & \sigma_Y^2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ &= \sigma_X^2 + 2\,\mathrm{Cov}(X, Y) + \sigma_Y^2,\end{aligned}$$

which you should recognize.

Example 27: *The bivariate normal with independent components*

> *Find the p.d.f. of the bivariate normal* $\mathbf{X} = (X_1, X_2)^T$ *in the case where* $\mathrm{Cov}(X_1, X_2) = 0$.

From Definition 6.4, the general bivariate normal distribution $\mathbf{X}$, with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ has joint probability density function

$$f_{X_1,X_2}(x_1, x_2) = \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}} \exp\left(-\frac{\sigma_2^2(x_1-\mu_1)^2 - 2\sigma_{12}(x_1-\mu_1)(x_2-\mu_2) + \sigma_1^2(x_2-\mu_2)^2}{2(\sigma_1^2\sigma_2^2 - \sigma_{12}^2)}\right)$$

If we assume $\mathrm{Cov}(X_1, X_2) = \sigma_{12} = \sigma_{21} = 0$, then the p.d.f. simplifies to

$$\begin{aligned} f_{X_1,X_2}(x_1, x_2) &= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma_1^2} - \frac{(x_1-\mu_2)^2}{2\sigma_2^2}\right) \\ &= \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}\right) \times \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{(x_2-\mu_2)^2}{2\sigma_2^2}\right) \\ &= f_{X_1}(x_1) f_{X_2}(x_2). \end{aligned} \tag{4}$$

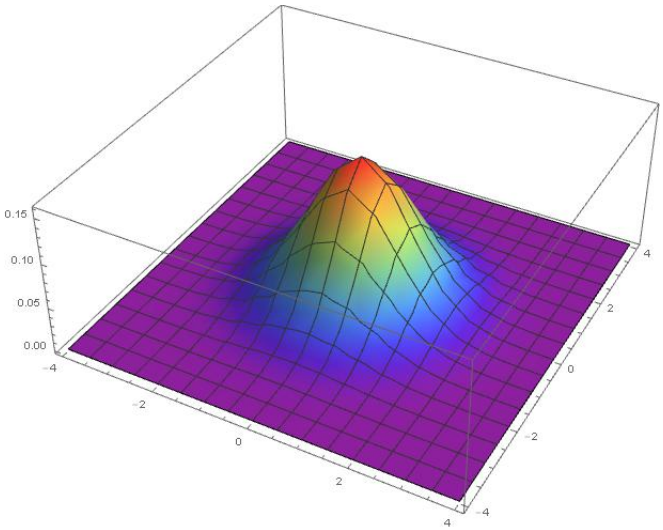
Here, in the final line we see factorize $f_{X_1,X_2}(x_1, x_2)$, into the product of the p.d.f. of the $N(\mu_1, \sigma_1^2)$ random variable X_1 and the p.d.f. of the $N(\mu_2, \sigma_2^2)$ random variables X_2. Therefore, in this case X_1 and X_2 are independent.

Note that, setting $\mu_1 = \mu_2 = 0$ and $\sigma_1 = \sigma_2 = 1$, we recover (6.1).

We have shown above that if $\mathrm{Cov}(X_1, X_2) = 0$ *then* X_1 *and* X_2 *are independent. If* X_1 *and* X_2 *are independent then it is automatic that* $\mathrm{Cov}(X_1, X_2) = 0$. *Hence:* X_1 *and* X_2 *are independent if and only if* $\mathrm{Cov}(X_1, X_2) = 0$. *We will record this fact as Lemma 6.8.*

Example 28: *Plotting the p.d.f. of the bivariate normal.*

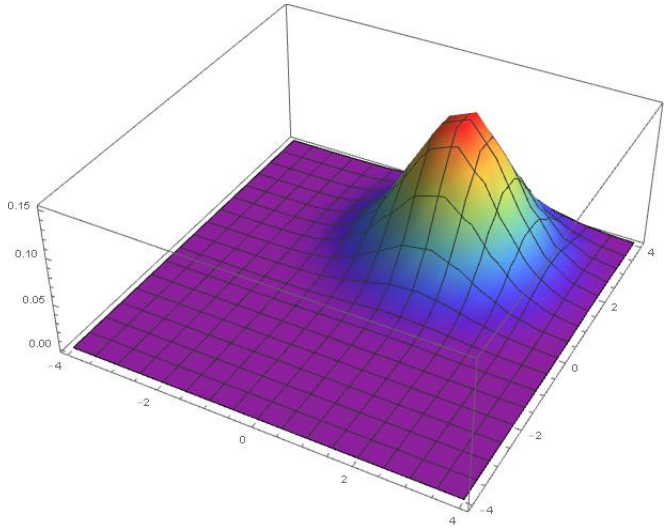
The pdf of a bivariate normal is a 'bell curve':



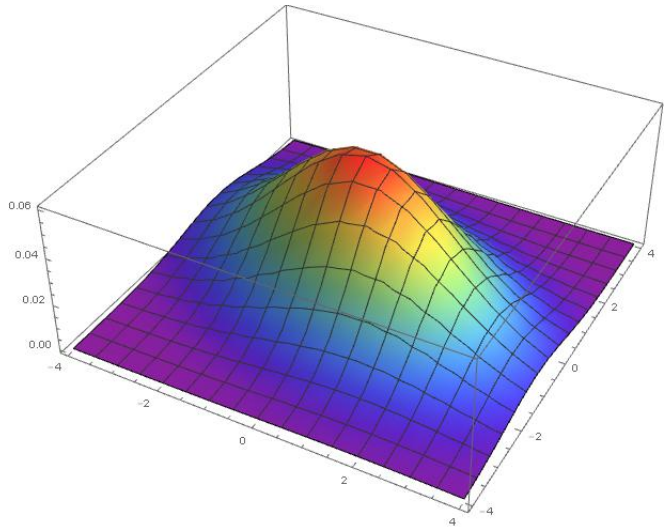
This example is the standard bivariate normal $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = (0, 0)$ and $\boldsymbol{\Sigma} = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$. It was generated in Mathematica with the code (all one line)

```
Plot3D[1/(2Pi) E^(-(x^2 + y^2)/2), {x, -4, 4}, {y, -4, 4}, PlotRange -> All,
ColorFunction -> (ColorData["Rainbow"][#3] &)]
```

Changing $\boldsymbol{\mu}$ alters the positive of the center of the bell, without changing the shape of the curve. For example, taking $\boldsymbol{\mu} = (1, 2)$ and $\boldsymbol{\Sigma} = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ gives



Changing $\boldsymbol{\Sigma}$ afters the shape of the bell. For example, taking $\boldsymbol{\mu} = (0, 0)$ and $\boldsymbol{\Sigma} = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 4 \end{smallmatrix}\right)$ gives



Changing both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ together results in a bell curve that is both translated and reshaped.

Example 29: *Marginal distributions of the bivariate normal, and their covariance.*

$>$ *Let* $\mathbf{X} = (X_1, X_2)^T$ *have distribution* $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *where* $\boldsymbol{\mu} = \left(\begin{smallmatrix} 1 \\ 3 \end{smallmatrix}\right)$ *and* $\boldsymbol{\Sigma} = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 3 \end{smallmatrix}\right)$. *Write down the marginal distributions of* X_1 *and* X_2.

From Lemma 6.7 we know that X_1 and X_2 are both (univariate) normals. We can read their means and covariances off from the mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. We have $X_1 \sim N(\mu_1, \sigma_{11})$, so $X_1 \sim N(1, 2)$, and also $X_2 \sim N(\mu_2, \sigma_{22})$ so $X_2 \sim N(3, 3)$.

> *Find* $\text{Cov}(X_1, X_2)$ *and* $\rho(X_1, X_2)$*. Are* X_1 *and* X_2 *independent?*

From the covariance matrix, $\text{Cov}(X_1, X_2) = 1$. Hence,

$$\rho(X_1, X_2) = \frac{\text{Cov}(X_1, X_2)}{\sqrt{\text{Var}(X_1)\,\text{Var}(X_2)}} = \frac{1}{\sqrt{2 \cdot 3}} = \frac{1}{\sqrt{6}}.$$

Clearly, we have $\text{Cov}(X_1, X_2) \neq 0$ so X_1 and X_2 are not independent.

Example 30: *Conditional distributions for bivariate normal*

> *Let* $a \in \mathbb{R}$ *and let* $X \sim N_2(\boldsymbol{\mu}, \Sigma)$ *where* $\boldsymbol{\mu}$ *and* Σ *are as in Example 31. Find the conditional distribution of* X_2 *given* $X_1 = a$*.*

By Lemma 6.9, the conditional distribution of X_2 given $X_1 = a$ is a univariate normal with mean given by $\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1)$ and variance $(1 - \rho^2)\sigma_2^2$.

In this case, $\mu_1 = 1$, $\mu_2 = 2$, $\rho = 3/\sqrt{10}$, $\sigma_2 = \sqrt{10}$, $\sigma_1 = 1$, and $x_1 = a$. So, $\mu = 2 + 3(a - 1)$ and $\sigma^2 = (1 - \frac{9}{10})(10) = 1$. Hence, the conditional distribution of X_2 given $X_1 = a$ is $N(2 + 3(a - 1), 1)$.

Example 31: *Transformations of bivariate normal*

> *Let* $X \sim N_2(\boldsymbol{\mu}, \Sigma)$ *where* $\boldsymbol{\mu} = \left(\begin{smallmatrix} 1 \\ 2 \end{smallmatrix}\right)$ *and* $\boldsymbol{\Sigma} = \left(\begin{smallmatrix} 1 & 3 \\ 3 & 10 \end{smallmatrix}\right)$*. Let*
>
> $$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_1 + 2X_2 - 1 \\ X_2 - 1 \end{pmatrix}$$
>
> *Find the distribution of* $\mathbf{Y} = (Y_1, Y_2)^T$*.*

We can write $\mathbf{Y}$ as an affine transformation of $\mathbf{X}$, that is

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b} = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} + \begin{pmatrix} -1 \\ -1 \end{pmatrix}.$$

The matrix $\mathbf{A} = \left(\begin{smallmatrix} 1 & 2 \\ 0 & 1 \end{smallmatrix}\right)$ is a non-singular 2×2 matrix, so by Lemma 6.10, $\mathbf{Y}$ is a bivariate normal.

Therefore, if we can find the mean vector and covariance matrix of $\mathbf{Y}$, we know the distribution of $\mathbf{Y}$.

$$\mathbb{E}[\mathbf{Y}] = \mathbf{A}\mathbb{E}[\mathbf{X}] + \mathbf{b} = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 4 \\ 1 \end{pmatrix},$$

and

$$\text{Cov}(\mathbf{Y}) = \mathbf{A}\,\text{Cov}(\mathbf{X})\mathbf{A}^T = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 3 \\ 3 & 10 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} 53 & 23 \\ 23 & 10 \end{pmatrix}.$$

So, the distribution of $\mathbf{Y}$ is

$$Y \sim N_2\left[\begin{pmatrix} 4 \\ 1 \end{pmatrix}, \begin{pmatrix} 53 & 23 \\ 23 & 10 \end{pmatrix}\right].$$

Example 32: *Affine transformation of a three dimensional normal distribution.*

> *Suppose*

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim N_3 \left[\begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 4 & 2 & 0 \\ 2 & 9 & 1 \\ 0 & 1 & 4 \end{pmatrix} \right].$$

Find the joint distribution of $\mathbf{Y} = (Y_1, Y_2)^T$ *where* $Y_1 = X_2 - X_1$ *and* $Y_2 = X_1 + X_2 + X_3$.

We can write,

$$\mathbf{Y} = \mathbf{AX} = \begin{pmatrix} -1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix},$$

so

$$\mathbb{E}[\mathbf{Y}] = \mathbf{A}\mathbb{E}[\mathbf{X}] = \begin{pmatrix} -1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 4 \end{pmatrix},$$

and

$$\begin{aligned} \operatorname{Cov}(\mathbf{Y}) &= \mathbf{A} \operatorname{Cov}(\mathbf{X}) \mathbf{A}^T \\ &= \begin{pmatrix} -1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 4 & 2 & 0 \\ 2 & 9 & 1 \\ 0 & 1 & 4 \end{pmatrix} \begin{pmatrix} -1 & 1 \\ 1 & 1 \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} -1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} -1 & 6 \\ 7 & 12 \\ 1 & 5 \end{pmatrix} \\ &= \begin{pmatrix} 9 & 6 \\ 6 & 23 \end{pmatrix}. \end{aligned}$$

It is not hard to see that $\mathbf{A}$ is an onto transformation, so $\mathbf{Y}$ has a bivariate normal distribution (here we use the multivariate equivalent of Lemma 6.10). Hence,

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N_2 \left[\begin{pmatrix} 1 \\ 4 \end{pmatrix}, \begin{pmatrix} 9 & 6 \\ 6 & 23 \end{pmatrix} \right].$$

> *Find* $\rho(X_1, X_3)$. *Are* X_1 *and* X_3 *independent?*

From the covariance matrix of $\mathbf{X}$, we can read off

$$\rho(X_1, X_3) = \frac{\operatorname{Cov}(X_1, X_3)}{\sqrt{\operatorname{Var}(X_1) \operatorname{Var}(X_3)}} = \frac{0}{\sqrt{4 \cdot 4}} = 0.$$

Since X_1 and X_3 are components of a multivariable normal distribution, and $\operatorname{Cov}(X_1, X_3) = 0$, by (the three dimensional equivalent of) Lemma 6.8 X_1 and X_3 are independent.

Chapter 7

Example 33: *Maximising a function*

> *Find the value of θ which maximises $f(\theta) = \theta^5(1-\theta)$ on the range $\theta \in [0,1]$.*

First, we look for turning points. We have

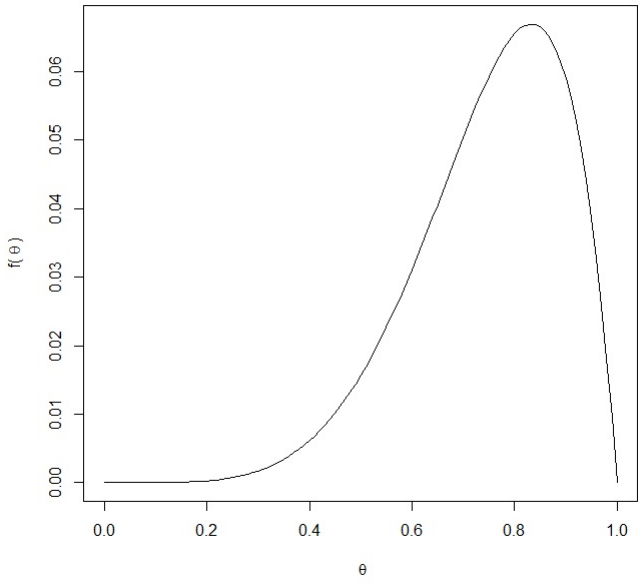
$$\begin{aligned} f'(\theta) &= 5\theta^4(1-\theta) + \theta^5(-1) \\ &= \theta^4(5-6\theta) \end{aligned}$$

So the turning points are at $\theta = 0$ and $\theta = \frac{5}{6}$. To see which ones are local maxima, we calculate the second derivative:

$$\begin{aligned} f''(\theta) &= 4\theta^3(5-6\theta) + \theta^4(-6) \\ &= \theta^3(20-30\theta). \end{aligned}$$

So, $f''(\frac{5}{6}) = (\frac{5}{6})^3(20-25) < 0$ and $\theta = \frac{5}{6}$ is a local maximum. Unfortunately, $f''(0) = 0$, so we don't know if $\theta = 0$ is a local maximum, minimum or inflection. However, we can check that $f(0) = 0$, so it doesn't matter which, we still have $f(0) < f(\frac{5}{6})$.

Hence, $\theta = \frac{5}{6}$ is the global maximiser.



Example 34: *Likelihood functions and maximum likelihood estimators*

> *Let X be a random variable with $Exp(\theta)$ distribution, where the parameter λ is unknown. Find the and sketch the likelihood function of X, given the data $x = 3$.*

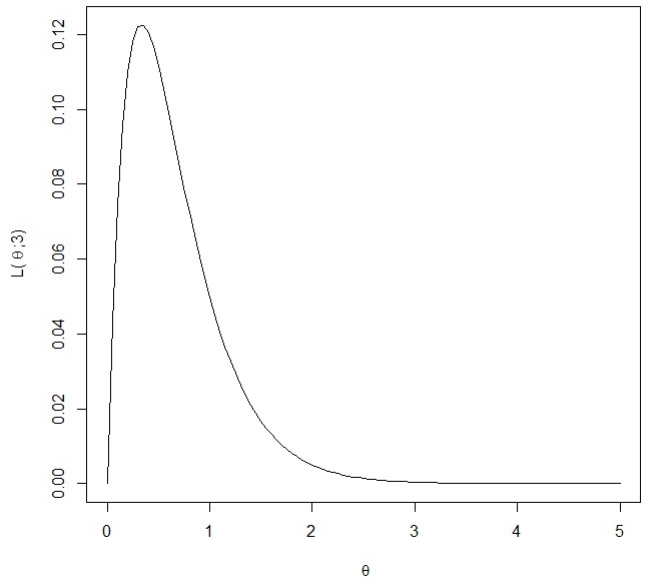
The likelihood function is

$$L(\theta; 3) = f_X(3; \theta) = \theta e^{-3\theta}$$

defined for all $\theta \in \Theta = (0, \infty)$. We can plot this in `R`, for $\theta \in (0, 20)$, with the command

```
curve(x*exp(-3x), from=0, to=5, xlab=~theta, ylab="L("~theta~";4))
```

(Note that we use `x` as the θ variable here because `R` hard-codes its use of x as a graph variable.) The result is



> *Given this data, find the likelihood of $\theta = \frac{1}{10}, \frac{1}{2}, 1, 2, 5$. Amongst these values of θ, which has the highest likelihood?*

The likelihoods are

$$\begin{aligned} L(\tfrac{1}{10};3) &= \tfrac{1}{10}e^{-\frac{3}{10}} &&\approx 0.07 \\ L(\tfrac{1}{2};3) &= \tfrac{1}{2}e^{-\frac{3}{2}} &&\approx 0.11 \\ L(1;3) &= 1e^{-3} &&\approx 0.05 \\ L(2;3) &= 2e^{-6} &&\approx 0.005 \\ L(5;3) &= 5e^{-15} &&\approx 1.5 \times 10^{-6} \end{aligned}$$

So, restricted to looking at these values, $\theta = \frac{1}{2}$ has the highest likelihood.

> *Find the maximum likelihood estimator of $\theta \in (0, \infty)$, based on the (single) data point $x = 3$.*

We need to find the value of $\theta \in \Theta$ which maximises $L(\theta; 3)$. We differentiate, to look for turning points, obtaining

$$\begin{aligned} \frac{dL}{d\theta} &= e^{-3\theta} - 3\theta e^{-3\theta} \\ &= e^{-3\theta}(1 - 3\theta). \end{aligned}$$

Hence, there is only one turning point, at $\theta = \frac{1}{3}$. We differentiate again, obtaining

$$\begin{aligned}\frac{d^2L}{d\theta^2} &= -3e^{-3\theta}(1-3\theta) + e^{-3\theta}(-3) \\ &= e^{-3\theta}(-6+9\theta)\end{aligned}$$

At $\theta = \frac{1}{3}$, we have $\frac{d^2L}{d\theta^2} = e^{-1}(-6+3) < 0$, so the turning point at $\theta = \frac{1}{3}$ is a local maximum. Since it is the only turning point, it is also the global maximum. Hence, the maximum likelihood estimator of θ is $\hat{\theta} = \frac{1}{3}$.

Example 35: *Models, parameters and data (aerosols).*

> *The 'particle size distribution' of an aerosol is the distribution of the diameter of aerosol particles within a typical region of air. The term is also used for particles within a powder, or suspended in a fluid.*
>
> *In many situations, the particle size distribution is modelled using the log-normal distribution. It is typically reasonable to assume that the diameters of particles are independent. Assuming this model, find the joint probability density function of the diameters observed in a sample of n particles, and state the parameters of the model.*

Recall that the p.d.f. of the log-normal distribution is

$$f_Y(y) = \begin{cases} \frac{1}{y\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log y - \mu)^2}{2\sigma^2}\right) & \text{if } y \in (0,\infty) \\ 0 & \text{otherwise.} \end{cases}$$

The parameters of this distribution, and hence also the parameters of our model, are $\mu \in \mathbb{R}$ and $\sigma \in (0,\infty)$. Since the diameters of particles are assumed to be independent, the joint probability density function of $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$, where Y_i is the diameter of the i^{th} particle, is

$$\begin{aligned} f_{\mathbf{Y}}(y_1,\ldots,y_n) &= \prod_{i=1}^{n} f_{Y_i}(y_i) \\ &= \begin{cases} \frac{1}{(2\pi\sigma^2)^{n/2}} \frac{1}{y_1 y_2 \ldots y_n} \exp\left(-\sum_{i=1}^{n} \frac{(\log y_i - \mu)^2}{2\sigma^2}\right) & \text{if } y_i > 0 \text{ for all } i \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Note that, if one (or more) of the y_i is less than or equal to zero then $f_{Y_i}(y_i) = 0$, which means that also $f_{\mathbf{Y}}(y_1,\ldots,y_n) = 0$.

Example 36: *Maximum likelihood estimation with i.i.d. data.*

> *Let $X \sim Bern(\theta)$, where θ is an unknown parameter. Suppose that we have 3 independent samples of X, which are*
>
> $$\mathbf{x} = \{0, 1, 1\}.$$
>
> *Find the likelihood function of θ, given this data.*

The probability function of a single $Bern(\theta)$ random variable is

$$f_X(x;\theta) = \begin{cases} \theta & \text{if } x = 1 \\ 1-\theta & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases}$$

Since our three samples are independent, we model $\mathbf{x}$ as a sample from the joint distribution $\mathbf{X} = (X_1, X_2, X_3)$, where

$$f_{\mathbf{X}}(\mathbf{x};\theta) = \prod_{i=1}^{3} f_{X_i}(x_i;\theta)$$

and f_{X_i} is the p.d.f. of a single $Bern(\theta)$ random variable. Since f_{X_i} has several cases, it would be unhelpful to try and expand out this formula before we put in values for the x_i. Our likelihood function is therefore

$$\begin{aligned} L(\theta;\mathbf{x}) &= f_{X_1}(0;\theta)\, f_{X_2}(1;\theta)\, f_{X_3}(1;\theta) \\ &= (1-\theta)\theta\theta \\ &= \theta^2 - \theta^3. \end{aligned}$$

The range of values that the parameter θ can take is $\Theta = [0,1]$.

> *Find the maximum likelihood estimator of θ, given the data $\mathbf{x}$.*

We seek to maximize $L(\theta;\mathbf{x})$ for $\theta \in [0,1]$. Differentiating once,

$$\frac{dL}{d\theta} = 2\theta - 3\theta^2 = \theta(2-3\theta)$$

so the turning points are at $\theta = 0$ and $\theta = \frac{2}{3}$. Differentiating again,

$$\frac{dL}{d\theta} = 2 - 6\theta$$

which gives $\frac{dL}{d\theta}\big|_{\theta=0} = 2$ and $\frac{dL}{d\theta}\big|_{\theta=2/3} = 2 - 4 = -2$. Hence, $\theta = 0$ is a local minimum and $\theta = \frac{2}{3}$ is a local maximum, so $\theta = \frac{2}{3}$ maximises $L(\theta;\mathbf{x})$ over $\theta \in [0,1]$. The maximum likelihood estimator of θ is therefore

$$\hat{\theta} = \frac{2}{3}.$$

This is, hopefully, reassuring. The number of 1s in our sample of 3 was 2, so (using independence) $\theta = \frac{2}{3}$ seems like a good guess. See **Q7.10** for a much more general case of this example.

Example 37: *Maximum likelihood estimation (radioactive decay).*

> *Atoms of radioactive elements decay as time passes, meaning that any such atom will, at some point in time, suddenly break apart. This process is known as 'radioactive decay'.*

The time taken for a single atom of, say, carbon-15 to decay is usually modelled as an exponential random variable, with unknown parameter $\lambda \in (0,\infty)$. The parameter λ is known as the 'decay rate'. The times at which atoms decay are known to be independent.

Using this model, find the likelihood function for the time to decay of a sample of n carbon-15 atoms.

The decay time X_i of the i^{th} atom is exponential with parameter $\lambda \in (0, \infty)$, and therefore has p.d.f.

$$f_{X_i}(x_i; \lambda) = \begin{cases} \lambda e^{-\lambda x_i} & \text{if } x_i > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Since each atom decays independently, the joint distribution of $\mathbf{X} = (X_i)_{i=1}^n$ is

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}; \lambda) = \prod_{i=1}^n f_{X_i}(x_i; \lambda) &= \begin{cases} \prod_{i=1}^n \lambda e^{-\lambda x_i} & \text{if } x_i > 0 \text{ for all } i \\ 0 & \text{otherwise.} \end{cases} \\ &= \begin{cases} \lambda^n \exp\left(-\lambda \sum_{i=1}^n x_i\right) & \text{if } x_i > 0 \text{ for all } i \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Therefore, the likelihood function is

$$L(\lambda; \mathbf{x}) = \begin{cases} \lambda^n \exp\left(-\lambda \sum_{i=1}^n x_i\right) & \text{if } x_i > 0 \text{ for all } i \\ 0 & \text{otherwise.} \end{cases}$$

The range of possible values of the parameter λ is $\Theta = (0, \infty)$.

> *Suppose that we have sampled the decay times of 15 carbon-15 atoms (in seconds, accurate to two decimal places), and found them to be*

$$\mathbf{x} = \{0.50, 2.19, 0.88, 4.06, 9.75, 2.62, 0.13, 2.70, 0.03, 0.28, 4.15, 9.52, 2.67, 3.79, 4.31\}.$$

Find the maximum likelihood estimator of λ, based on this data.

Given this data, for which $\sum_{i=1}^{15} x_i = 47.58$, our likelihood function is

$$L(\lambda; \mathbf{x}) = \lambda^{15} e^{-47.58\lambda}.$$

Differentiating, we have

$$\begin{aligned} \frac{dL}{d\lambda} &= 15\lambda^{14} e^{-47.58\lambda} - 47.58\lambda^{15} e^{-47.58\lambda} \\ &= \lambda^{14}(15 - 47.58\lambda) e^{-47.58\lambda} \end{aligned}$$

which is zero only when $\lambda = 0$ or $\lambda = 15/47.58 \approx 0.32$. Since $\lambda = 0$ is outside of the range $\Theta = (0, \infty)$ of possible parameter values, the only turning point of interest is $\lambda = 15/47.58$.

Differentiating again (with the details left to you), we end up with

$$\begin{aligned} \frac{d^2L}{d\lambda^2} &= (210\lambda^{13} - 1427.4\lambda^{14} + 2263.86\lambda^{15}) e^{-47.58\lambda} \\ &= \lambda^{13}\left(210 - 1427.4\lambda + 2263.86\lambda^2\right) e^{-47.58\lambda} \end{aligned}$$

Evaluating at our turning point gives

$$\frac{d^2L}{d\lambda^2}\Big|_{\lambda=15/47.58} = \left(\frac{15}{47.58}\right)^{13}(-14.9996)\,e^{-15} < 0.$$

So, our turning point is a local maximum. Since there are no other turning points (within the allowable range) our turning point is the global maximum. Hence, the maximum likelihood estimator of λ, given our data $\mathbf{x}$, is

$$\hat{\lambda} = \frac{15}{47.58} \approx 0.32.$$

In reality, physicists are able to collect vastly more data than $n = 15$, but even with 15 data points we are not far away from the true value of λ, which is $\lambda \approx 0.283033$. Of course, by 'true' value here we mean the value that has been discovered experimentally, with the help of statistical inference.

So-called 'carbon dating' typically uses carbon-14, which has a much slower decay rate of approximately 1.21×10^{-4}. Carbon-14 is present in many living organisms and, crucially, the proportion of carbon in living organisms that is carbon-14 is essentially the same for all living organisms. Once organisms die, the carbon-14 radioactively decays. The key idea behind carbon dating is that, by measuring the concentration of carbon-14 within a fossil, scientists can estimate how long ago that fossil lived. To do so, a highly accurate estimate of the decay rate of carbon-14 is needed.

Example 38: *Maximum likelihood estimation via log-likelihood (mutations in DNA).*

> When organisms reproduce, the DNA (or RNA) of the offspring is a combination of the DNA of its (one, or two) parents. Additionally, the DNA of the offspring contains a small number of locations in which it differs from its parent(s). These locations are called 'mutations'.

The number of mutations per unit length of DNA is typically[1] modelled using a Poisson distribution, with an unknown parameter $\theta \in (0, \infty)$. The numbers of mutations found in disjoint sections of DNA are independent.

Using this model, find the likelihood function for the number of mutations present in a sample of n (disjoint) strands of DNA, each of which has unit length.

Let X_i be the number of mutations in the i^{th} strand of DNA. So, under our model,

$$f_{X_i}(x_i;\theta) = \frac{e^{-\theta}\theta^{x_i}}{(x_i)!}$$

for $x_i \in \{0, 1, 2, \ldots\}$, and $f_{X_i}(x_i) = 0$ if $x_i \notin \mathbb{N} \cup \{0\}$. Since we assume the (X_i) are independent, the joint distribution of $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ has probability function

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{(x_i)!} \\ &= \frac{1}{(x_1)!(x_2)!\ldots(x_n)!} e^{-n\theta}\theta^{\sum_1^n x_i} \end{aligned}$$

[1] Actually, the biological details here are rather complicated, and we omit discussion of them.

provided all $x_i \in \mathbb{N} \cup \{0\}$, and zero otherwise. Therefore, our likelihood function is

$$L(\theta; \mathbf{x}) = \frac{1}{(x_1)!(x_2)! \dots (x_n)!} e^{-n\theta} \theta^{\sum_1^n x_i}.$$

The range of possible values for θ is $\Theta = (0, \infty)$.

> *Let $\mathbf{x}$ be a vector of data, where x_i is the number of mutations observed in a (distinct) unit length segment of DNA. Suppose that at least one of the x_i is non-zero.*
>
> *Find the corresponding log-likelihood function, and hence find the maximum likelihood estimator of θ.*

The log-likelihood function is $\ell(\theta; \mathbf{x}) = \log L(\theta; \mathbf{x})$, so

$$\begin{aligned} \log L(\theta, \mathbf{x}) &= \log \left(\frac{1}{(x_1)!(x_2)! \dots (x_n)!} e^{-n\theta} \theta^{\sum_1^n x_i} \right) \\ &= \sum_{i=1}^{n} (-\log(x_i)!) - n\theta + (\log \theta) \sum_{i=1}^{n} x_i. \end{aligned}$$

We now look to maximise $l(\theta; \mathbf{x})$, over $\theta \in (0, \infty)$. Differentiating, we obtain

$$\frac{d\ell}{d\theta} = -n + \frac{1}{\theta} \sum_{i=1}^{n} x_i.$$

Note that this is much simpler than what we'd get if we differentiated $L(\theta; \mathbf{x})$. So, the only turning point of $\ell(\theta, \mathbf{x})$ is at $\theta = \frac{1}{n} \sum_{i=1}^{n} x_i$. Differentiating again, we have

$$\frac{d^2\ell}{d\theta^2} = -\frac{1}{\theta^2} \sum_{i=1}^{n} x_i.$$

Since our x_i are counting the occurrences of mutations, $x_i \geq 0$, and since at least one is non-zero we have $\frac{d^2 l}{d\theta^2} < 0$ (for all θ). Hence, our turning point is a maximum and, since it is the only maximum, is also the global maximum. Therefore, the maximum likelihood estimator of θ is

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

> *Mutations rates were measured, for 11 HIV patients, and there were found to be*
>
> $$\mathbf{x} = \Big\{ 19,\ 16,\ 37,\ 28,\ 24,\ 34,\ 37,\ 126,\ 32,\ 48,\ 45 \Big\}$$
>
> *mutations per 10^4 possible locations (i.e. 'per unit length'). This data comes from the article Cuevas et al. (2015) [2].*
>
> *Assuming the model suggested above, calculate the maximum likelihood estimator of the mutation rate of HIV.*

The data has

$$\bar{x} = \frac{1}{11} \sum_{i=1}^{11} x_i = \frac{446}{11} \approx 41$$

so we conclude that the maximum likelihood estimator of the mutation rate θ, given this data, is $\hat{\theta} = \frac{446}{11}$.

[2] http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002251

Example 39: *Maximum likelihood estimation via log-likelihood (spectrometry).*

> Using a mass spectrometer, it is possible to measure the mass[3] of individual molecules. For example, it is possible to measure the masses of individual amino acid molecules.

A sample of 15 amino acid molecules, which are all known to be of the same type (and therefore, the same mass), were reported to have masses

$$\mathbf{x} = \{65.76,\ 140.40,\ 94.02,\ 32.23,\ 115.00,\ 4.77,\ 116.00,\ 86.41,\ 91.14,\ 66.27,\ 91.00,\ 144.7.\ 39.33,\ 58.90\}.$$

It is known that these molecules are either Alanine, which has mass 71.0, or Leucine, which has mass 113.1. Given a molecule of mass θ, the spectrometer is known to report its mass as $X \sim N(\theta, 35^2)$, independently for each molecule.

Using this model, and the data above, find the likelihoods of Alanine and Leucine. Specify which of these has the greatest the likelihood.

Our model, for the reported mass X of a single molecule with (real) weight θ, is $X \sim N(0, 35^2)$. Therefore, $X_i \sim N(\theta, 30^2)$ and the p.d.f. of a single data point is

$$f_{X_i}(x_i) = \frac{1}{\sqrt{2\pi}35} \exp\left(-\frac{(x_i-\theta)^2}{2\times 35^2}\right).$$

Therefore, the p.d.f. of the reported masses $\mathbf{X} = (X_1, \ldots, X_n)$ of n molecules is

$$f_{\mathbf{X}}(x) = \prod_{i=1}^{n} f_{X_i}(x_i) = \frac{1}{(2\pi)^{n/2}35^n} \exp\left(-\frac{1}{2450}\sum_{i=1}^{n}(x_i-\theta)^2\right).$$

We know that, in reality, θ must be one of only two different values; 71.0 (for Alanine) and 113.1 (for Leucine). Therefore, our likelihood function is

$$L(\theta; \mathbf{x}) = \frac{1}{(2\pi)^{n/2}35^n} \exp\left(-\frac{1}{2450}\sum_{i=1}^{n}(x_i-\theta)^2\right)$$

and the possible range of values for θ is the two point set $\Theta = \{71.0, 113.1\}$. We need to find out which of these two values maximises the likelihood.

Our data $\mathbf{x}$ contains $n = 15$ data points. A short calculation (use e.g. `R`) shows that

$$\frac{1}{2450}\sum_{i=1}^{15}(x_i-71.0)^2 \approx 12.70, \qquad \frac{1}{2450}\sum_{i=1}^{15}(x_i-113.1)^2 \approx 20.41.$$

and, therefore, that

$$L(71.0; \mathbf{x}) \approx 2.19\times 10^{-34}, \qquad L(113.1; \mathbf{x}) = 9.90\times 10^{-38}.$$

We conclude that $\theta = 71.0$ has (much) greater likelihood than $\theta = 113.1$, so we expect that the molecules sampled are Alanine.

[3] This is a simplification; in reality a mass spectrometer measure the mass to charge ratio of the molecule, but since the charges of molecule are already known, the mass can be inferred later. Atomic masses are measured in so-called 'atomic mass units'.

Note that, if we were to differentiate (as we did in other examples), we would find the maximiser θ for $L(\theta;\mathbf{x})$ *across the whole range* $\theta \in (-\infty, \infty)$, which turns out to be $\theta = 81.07$. This is not what we want here! The design of our experiment has meant that the range of possible values for θ is restricted to the two point set $\Theta = \{71.0, 113.1\}$. See **Q7.5** for the 'unrestricted' case.

Example 40: *Two parameter maximum likelihood estimation (rainfall).*

> *Find the maximum likelihood estimator of the parameter vector* $\boldsymbol{\theta} = (\mu, \sigma^2)$ *when the data* $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ *are modelled as i.i.d. samples from a normal distribution* $N(\mu, \sigma^2)$.

Our parameter vector is $\boldsymbol{\theta} = (\mu, \sigma^2)$, so let us write $v = \sigma^2$ to avoid confusion. As a result, we are interested in the parameters $\boldsymbol{\theta} = (\mu, v)$, and the range of possible values of $\boldsymbol{\theta}$ is $\Theta = \mathbb{R} \times (0, \infty)$.

The p.d.f. of the univariate normal distribution $N(\mu, v)$ is

$$f_X(x) = \frac{1}{\sqrt{2\pi v}} e^{-(x-\mu)^2/2v}.$$

Writing $\mathbf{X} = (X_1, \ldots, X_n)$, where the X_i are i.i.d. univariate $N(\mu, v)$ random variables, the likelihood function of $\mathbf{X}$ is

$$L(\boldsymbol{\theta};\mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi v)^{n/2}} \exp\left(-\frac{1}{2v}\sum_{i=1}^{n}(x_i - \mu)^2\right).$$

Therefore, the log likelihood is

$$\ell(\boldsymbol{\theta};\mathbf{x}) = -\frac{n}{2}\left(\log(2\pi) + \log(v)\right) - \frac{1}{2v}\sum_{i=1}^{n}(x_i - \mu)^2.$$

We now look to maximise $\ell(\boldsymbol{\theta};\mathbf{x})$ over $\boldsymbol{\theta} \in \Theta$. The partial derivatives are

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{v}\sum_{i=1}^{n}(x_i - \mu) = \frac{1}{v}\left(\sum_{i=1}^{n} x_i - n\mu\right)$$
$$\frac{\partial \ell}{\partial v} = -\frac{n}{2v} + \frac{1}{2v^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Solving $\frac{\partial \ell}{\partial \mu} = 0$ gives $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$. Solving $\frac{\partial \ell}{\partial v} = 0$ gives $v = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$. So both partial derivatives will be zero if and only if

$$\mu = \bar{x}, \qquad v = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{5}$$

This gives us the value of $\theta = (\mu, v)$ at the (single) turning point of ℓ.

Next, we use the Hessian matrix to check if this point is a local maximum. We have

$$\frac{\partial^2 \ell}{\partial \mu^2} = -\frac{n}{v}$$
$$\frac{\partial^2 \ell}{\partial \mu \partial v} = \frac{-1}{v^2}\left(\sum_{i=1}^{n} x_i - n\mu\right)$$
$$\frac{\partial^2 \ell}{\partial v^2} = \frac{n}{2v^2} - \frac{1}{v^3}\sum_{i=1}^{n}(x_i - \mu)^2$$

Evaluating these at our turning point, we get

$$\left.\frac{\partial^2 \ell}{\partial \mu^2}\right|_{(5)} = -\frac{n}{\hat{v}}$$
$$\left.\frac{\partial^2 \ell}{\partial \mu \partial v}\right|_{(5)} = \frac{-1}{v^2}\left(\sum_{i=1}^{n} x_i - n\bar{x}\right) = 0$$
$$\left.\frac{\partial^2 \ell}{\partial v^2}\right|_{(5)} = \frac{n}{2v^2} - \frac{1}{v^3}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n}{2v^2} - \frac{1}{v^3}n\hat{v} = \frac{-n}{2v^2}$$

so

$$H = \begin{pmatrix} -\frac{n}{v} & 0 \\ 0 & \frac{-n}{2v^2} \end{pmatrix}.$$

Since $-\frac{n}{v} < 0$ and $\det H = \frac{n^2}{2v^3} > 0$, our turning point (5) is a local maximum. Since it is the only turning point, it is also the global maximum. Hence, the MLE is

$$\hat{\mu} = \bar{x}$$
$$\hat{\sigma}^2 = \hat{v} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Note $\hat{\mu}$ is the sample mean, and $\hat{\sigma}^2$ is the (biased) sample variance.

> *For the years 1985-2015, the amount of rainfall (in milimeters) recorded as falling on Sheffield in December is as follows:*

{78.0, 142.3, 38.2, 36.0, 159.1, 136.0, 78.4, 67.4, 171.4, 103.9, 70.4, 98.2, 79.4, 57.9, 135.6, 118.0, 28.0, 129.8, 106.5, 46.3, 56.7, 114.0, 74.9, 52.8, 66.1, 18.8, 124.6, 136.0, 69.8, 102.0, 121.2}

This data comes from the historical climate data stored by the Met Office[4]*.*

Meteorologists often model the long run distribution of rainfall by a normal distribution (although in some cases the Gamma distribution is used). Assuming that we choose to model the amount of rainfall in Sheffield each December by a normal distribution, find the maximum likelihood estimators for μ and σ^2.

The data has $n = 30$, and

$$\bar{x} = \frac{1}{30}\sum_{i=1}^{30} \approx 93.9, \qquad \frac{1}{30}\sum_{i=1}^{30}(x_i - \bar{x})^2 \approx 1631.2 \approx 40.4^2$$

[4] http://www.metoffice.gov.uk/public/weather/climate-historic/

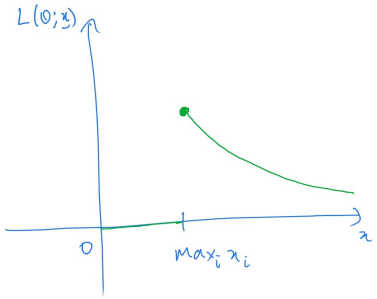
So we conclude that, according to our model, the maximum likelihood estimators are $\hat{\mu} \approx 93.9$ and $\hat{\sigma^2} \approx 40.4^2$, which means that Sheffield receives a $N(93.9, 40.4^2)$ quantity of rainfall, in millimetres, each December.

Example 41: *Maximum likelihood estimation for the uniform distribution*

> *Find the maximum likelihood estimator of the parameter θ when the data $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ are i.i.d. samples from a uniform distribution $U[0, \theta]$, with unknown parameter $\theta > 0$.*

Here the p.d.f. of X_i is $f(x) = \frac{1}{\theta}$ for $0 \leq x \leq \theta$ and zero otherwise. So the likelihood, for $\theta \in \Theta = \mathbb{R}^+$, is

$$L(\theta; \mathbf{x}) = \begin{cases} \frac{1}{\theta^n} & \text{if } \theta \geq x_i \text{ for all } i \\ 0 & \text{if } \theta < x_i \text{ for some } i \end{cases}$$
$$= \begin{cases} \frac{1}{\theta^n} & \text{if } \theta \geq \max_i x_i \\ 0 & \text{if } \theta < \max_i x_i. \end{cases}$$



Differentiating the likelihood, we see that $L(\theta; \mathbf{x})$ is decreasing (but positive) for $\theta > \max_i x_i$. For $\theta < \max_i x_i$ we know $L(\theta; \mathbf{x}) = 0$, so by looking at the graph, we can see that the maximum occurs at

$$\theta = \hat{\theta} = \max_{i=1,\ldots,n} x_i.$$

This is the MLE.

Example 42: *Interval estimation based on likelihood*

> *Suppose that we have i.i.d. data $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, for which each data point is modelled as a random sample from $N(\mu, \sigma^2)$ where μ is unknown and σ^2 is known. Find the k-likelihood region R_k for the parameter μ.*

First, we need to find the MLE $\hat{\mu}$ of μ. The likelihood function for our model is

$$L(\mu; \mathbf{x}) = \prod_{i=1}^{n} \phi(x_i; \mu) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right),$$

where the range of parameter values is all $\mu \in \mathbb{R}$. The log likelihood is

$$\ell(\mu; \mathbf{x}) = -\frac{n}{2} \left(\log(2\pi) + \log(\sigma^2) \right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2.$$

The usual process of maximisation (which is left for you and is a simplified case of Example 40) shows that the maximum likelihood estimator is the sample mean,

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Now we are ready to identify the k-likelihood region for μ. By definition, the k-likelihood region is

$$R_k = \{\mu \in \mathbb{R} \ : \ |l(\mu; \mathbf{x}) - l(\hat{\mu}; \mathbf{x})| \leq k\}.$$

So, $\mu \in R_k$ if and only if

$$\left| \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 \right| \leq k.$$

We can simplify this inequality, by noting that

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \mu)^2 - \sum_{i=1}^{n}(x_i - \hat{\mu})^2 &= \sum_{i=1}^{n} x_i^2 - 2x_i\mu + \mu^2 - x_i^2 + 2x_i\hat{\mu} - \hat{\mu}^2 \\
&= n\mu^2 - n\hat{\mu}^2 + 2(\hat{\mu} - \mu)\sum_{i=1}^{n} x_i \\
&= n\mu^2 - n\hat{\mu}^2 + 2(\hat{\mu} - \mu)n\hat{\mu} \\
&= n(\mu^2 + \hat{\mu}^2 - 2\mu\hat{\mu}) \\
&= n(\hat{\mu} - \mu)^2.
\end{aligned}$$

So, $\mu \in R_k$ if and only if

$$\frac{n}{2\sigma^2}|\hat{\mu} - \mu|^2 \leq k,$$

or in other words,

$$R_k = \left[\hat{\mu} - \sigma\sqrt{\frac{2k}{n}},\ \hat{\mu} + \sigma\sqrt{\frac{2k}{n}}\right].$$

Example 43: *Hypothesis tests based on likelihood*

> In Example 37, if we used a 2-likelihood test, would we accept the hypothesis that the radioactive decay of carbon-15 is equal to $\lambda = 0.27$?

We had found, given the data, that the likelihood function of θ was

$$L(\lambda; \mathbf{x}) = \lambda^{15} e^{-47.58\lambda}$$

and the maximum likelihood estimator of λ was $\hat{\lambda} \approx 0.32$. The 2-likelihood region for λ is the set

$$R_2 = \left\{\lambda > 0 \ : \ L(\lambda; \mathbf{x}) \geq e^{-2} L(\hat{\lambda}; \mathbf{x})\right\},$$

so $\lambda \in R_2$ if and only if

$$\lambda^{15} e^{-47.58\lambda} \geq e^{-2} L(0.32; \mathbf{x}) = 1.24 \times 10^{-15}.$$

Note that, unlike the previous example, we can't simplify this inequality and find a 'nice' form for the likelihood region.

Our hypothesis is that, in fact, $\lambda = 0.27$. Our 2-likelihood test will pass if $\lambda = 0.27$ is within the 2-likelihood region, and fail if not. We can evaluate (use e.g. `R`),

$$0.27^{15} e^{-47.58 \times 0.27} \approx 7.78 \times 10^{-15}$$

and note that $7.78 \times 10^{-15} \geq 1.24 \times 10^{-15}$. Hence $\lambda = 0.27$ is within the 2-likelihood region and we accept the hypothesis.